Classificação 2.0

Gabriel Fernandes Silva

Pré-Processamento

1420 IES

1335 ativas

85 extintas

#	Column	Non-Null Count	Dtype
0	Codigo da IES	1420 non-null	int64
1	Fim Lucrativo	1420 non-null	int64
2	Situacao	1420 non-null	int64
3	DIAS	1420 non-null	int64
4	Cursos Sem Ato 5 anos	1420 non-null	float64
5	Ultimo CI	1420 non-null	float64
6	Ultimo IGC	1420 non-null	float64
7	EAD_17	1420 non-null	int64
8	Variacao Matricula 16/17	1420 non-null	float64
9	Matriculas 17	1420 non-null	int64
10	% FIES	1420 non-null	float64
11	CURSOS	1420 non-null	int64
12	Saldo 2017	1420 non-null	float64
13	Variacao do Saldo 16/17	1420 non-null	float64

Balanceamento

Antes do Oversampling:

Extintas: 66 Ativas: 1070

Depois do Oversampling:

Extintas: 1070 Ativas: 1070

```
from sklearn.model_selection import train_test_split

colunas = ['Fim Lucrativo', 'DIAS', 'Cursos Sem Ato 5 anos', 'Ultimo CI',
           'Ultimo IGC', 'EAD_17', 'Variacao Matricula 16/17', 'Matriculas 17',
           '% FIES', 'CURSOS', 'Saldo 2017', 'Variacao do Saldo 16/17']

X = dados[colunas]
Y = dados['Situacao']

X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

```
smote = SMOTE()
X_train_balanceado, y_train_balanceado = smote.fit_resample( X_train, y_train )
```

Modelo 1

KNN

Separeção dos Dados

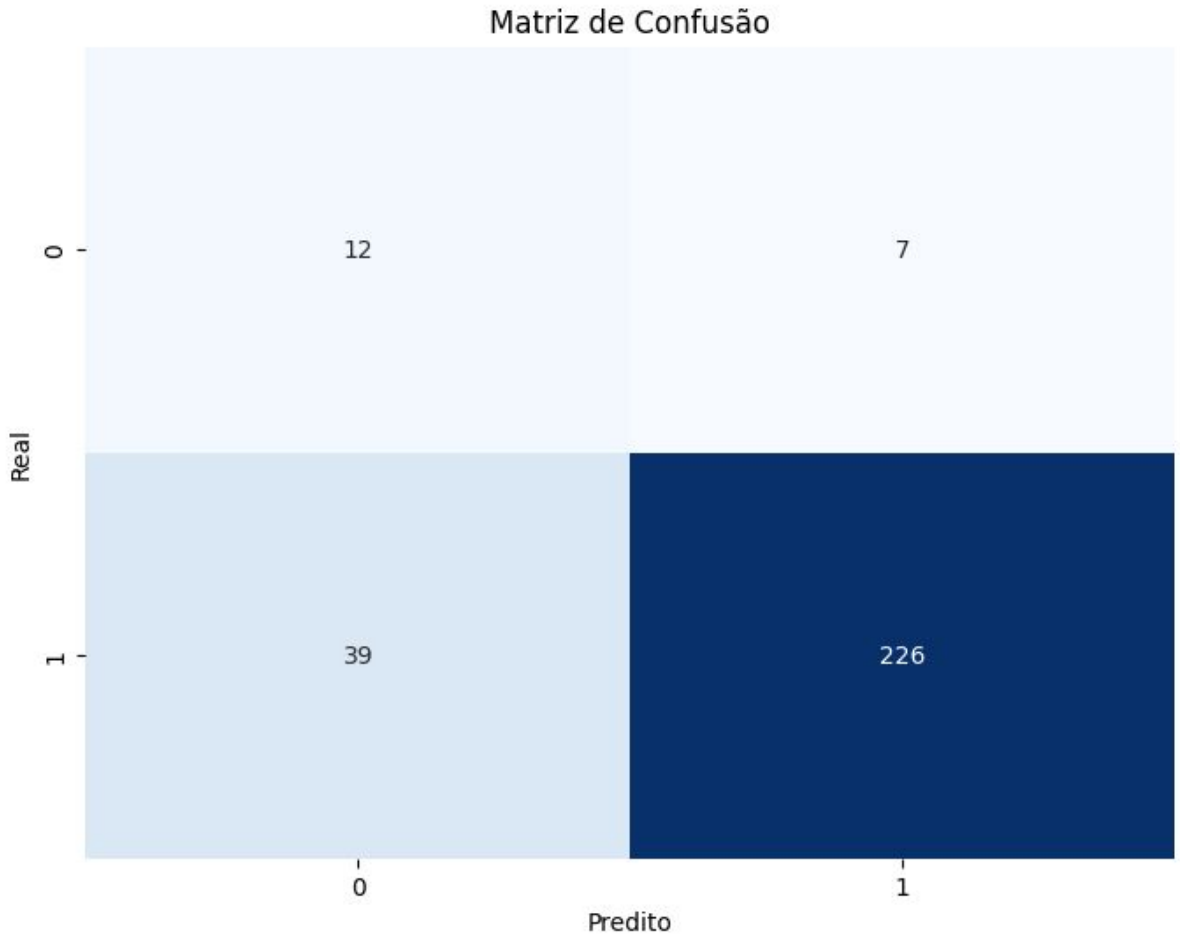
- 80% treino
- 20% teste

- n_neighbors = 3

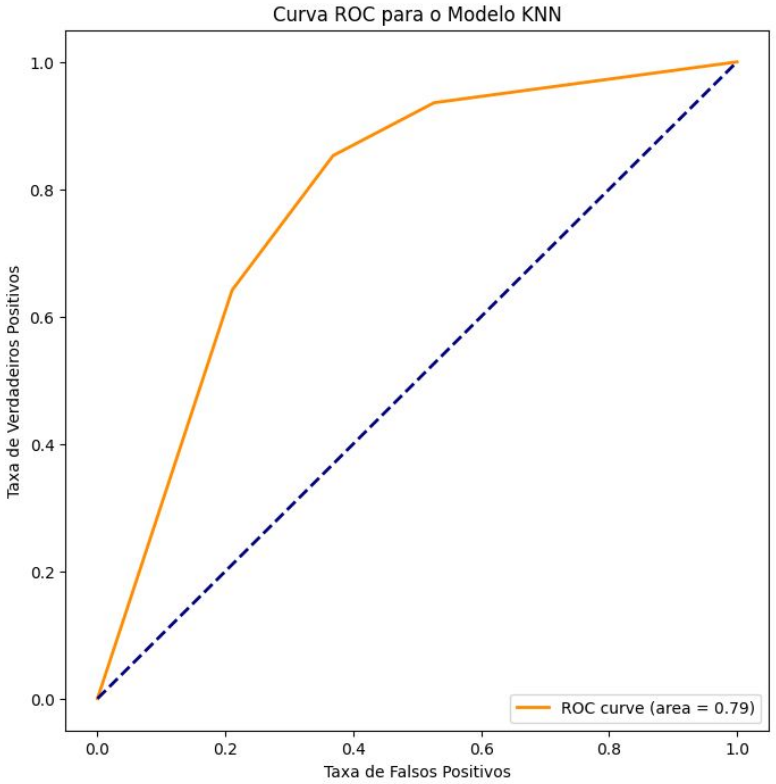
Métricas

	P	R	F1	S
Extinta	24%	63%	34%	19
Ativas	97%	85%	91%	265

Acurácia: 84%



Curva ROC


Curva ROC para o Modelo KNN
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.79)

Modelo 2

Árvore de Decisão

Separeção dos Dados

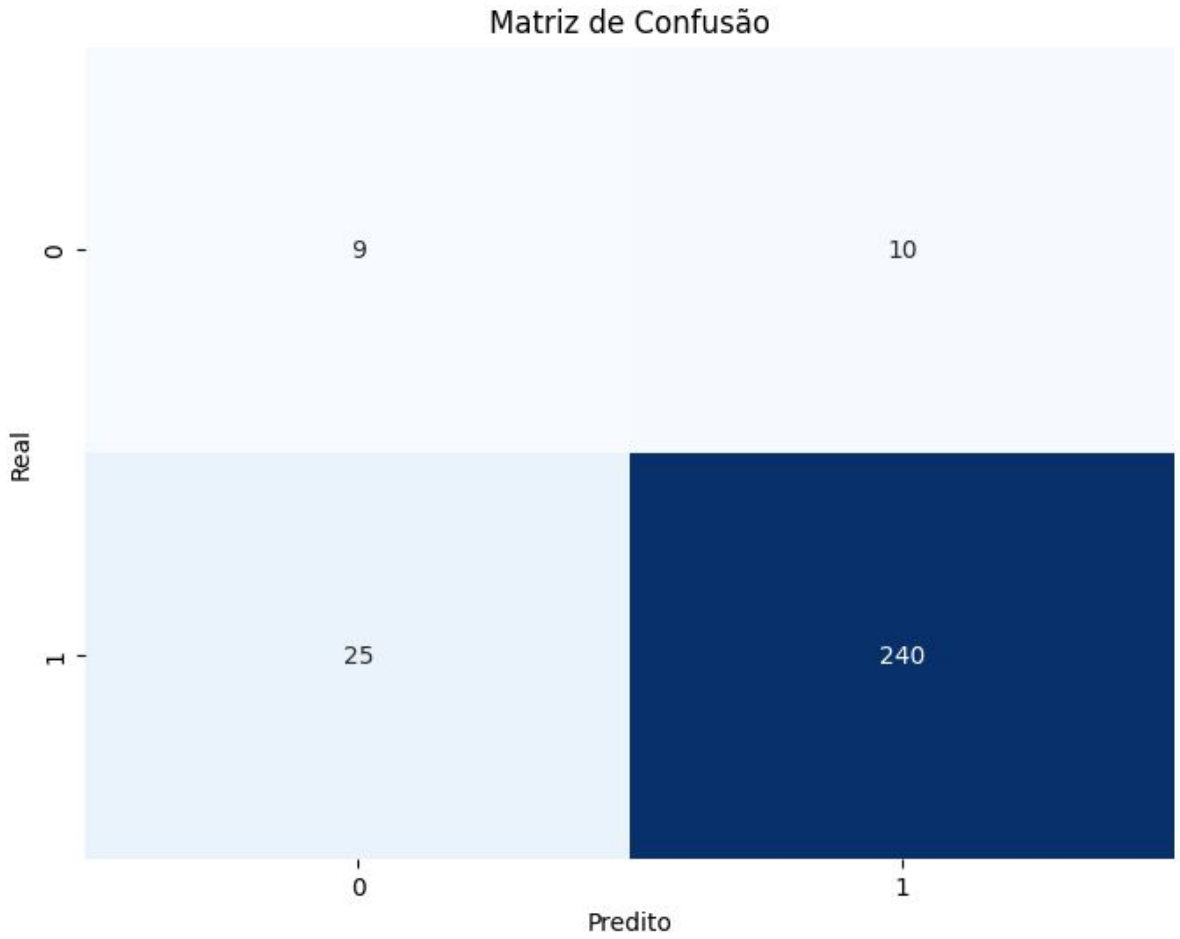
- 80% treino
- 20% teste

- Sem configuração de hiperparâmetros

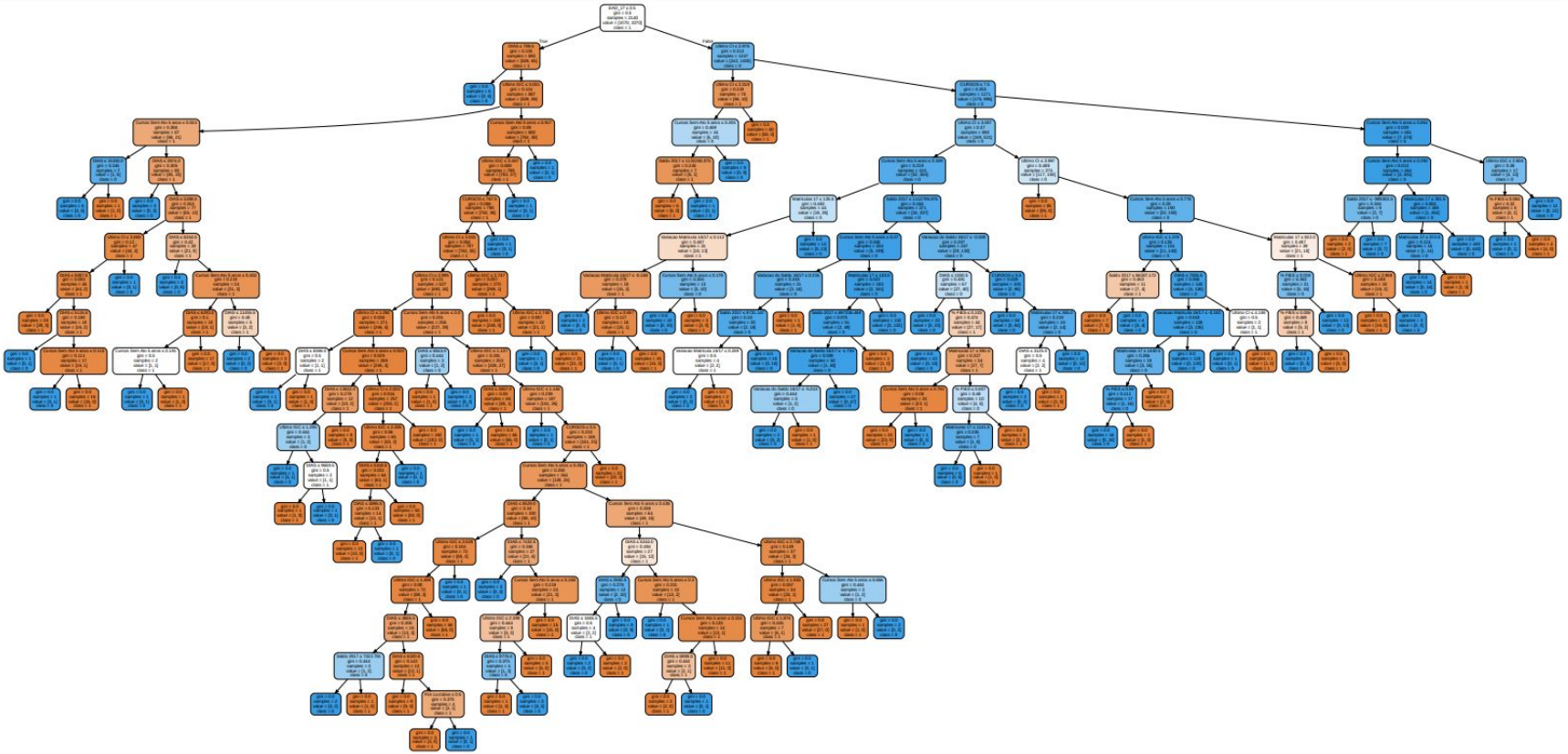
Métricas

	P	R	F1	S
Extinta	26%	47%	34%	19
Ativas	96%	91%	93%	265

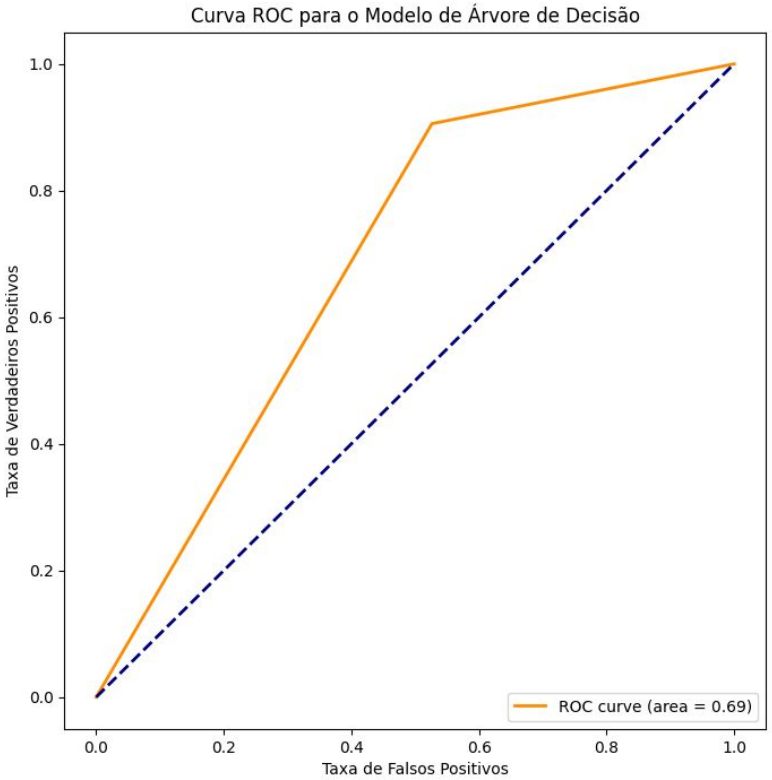
Acurácia: 88%



Visualização da árvore



Curva ROC


Curva ROC para o Modelo de Árvore de Decisão
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.69)

Separeção dos Dados

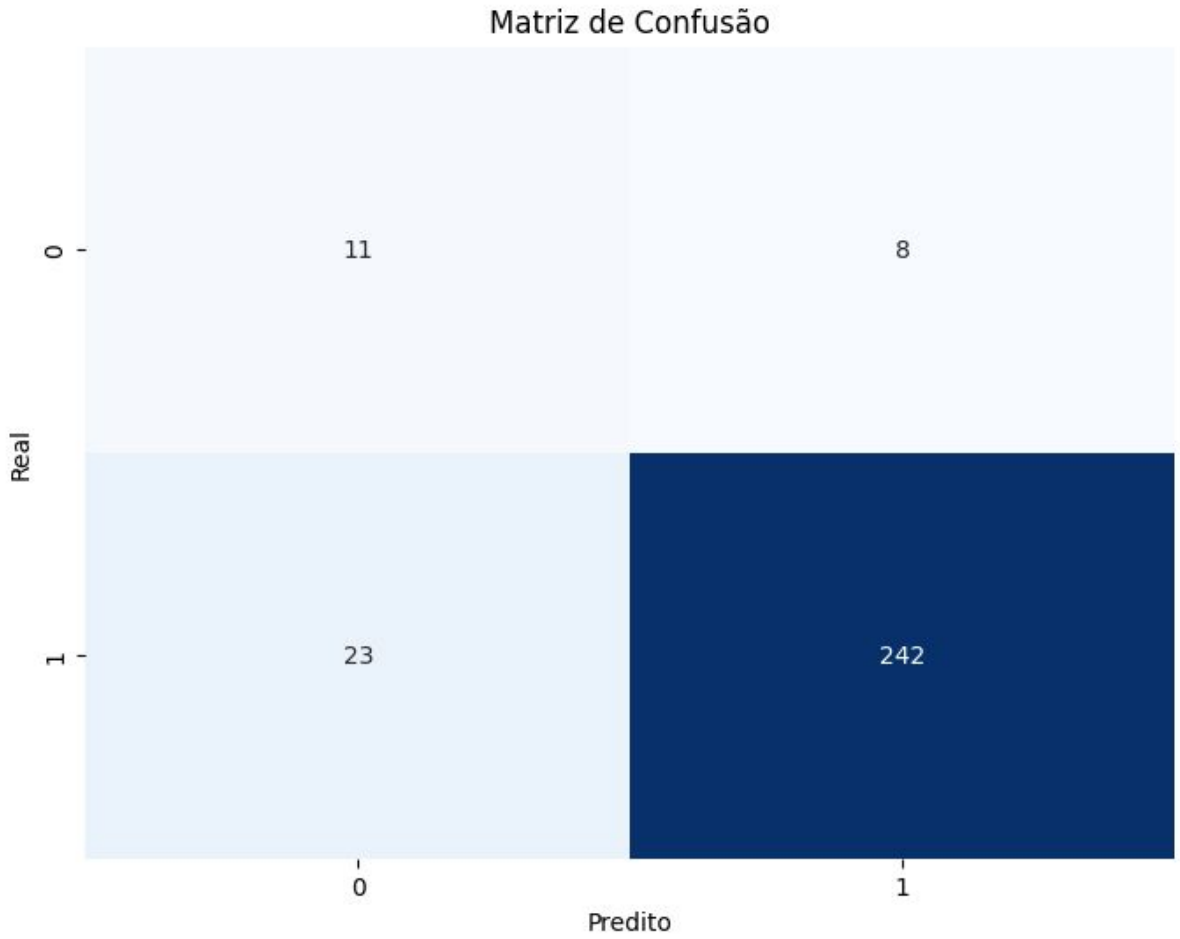
- 80% treino
- 20% teste

- max_depth = 5
- min_samples_split = 5
- min_samples_leaf = 5

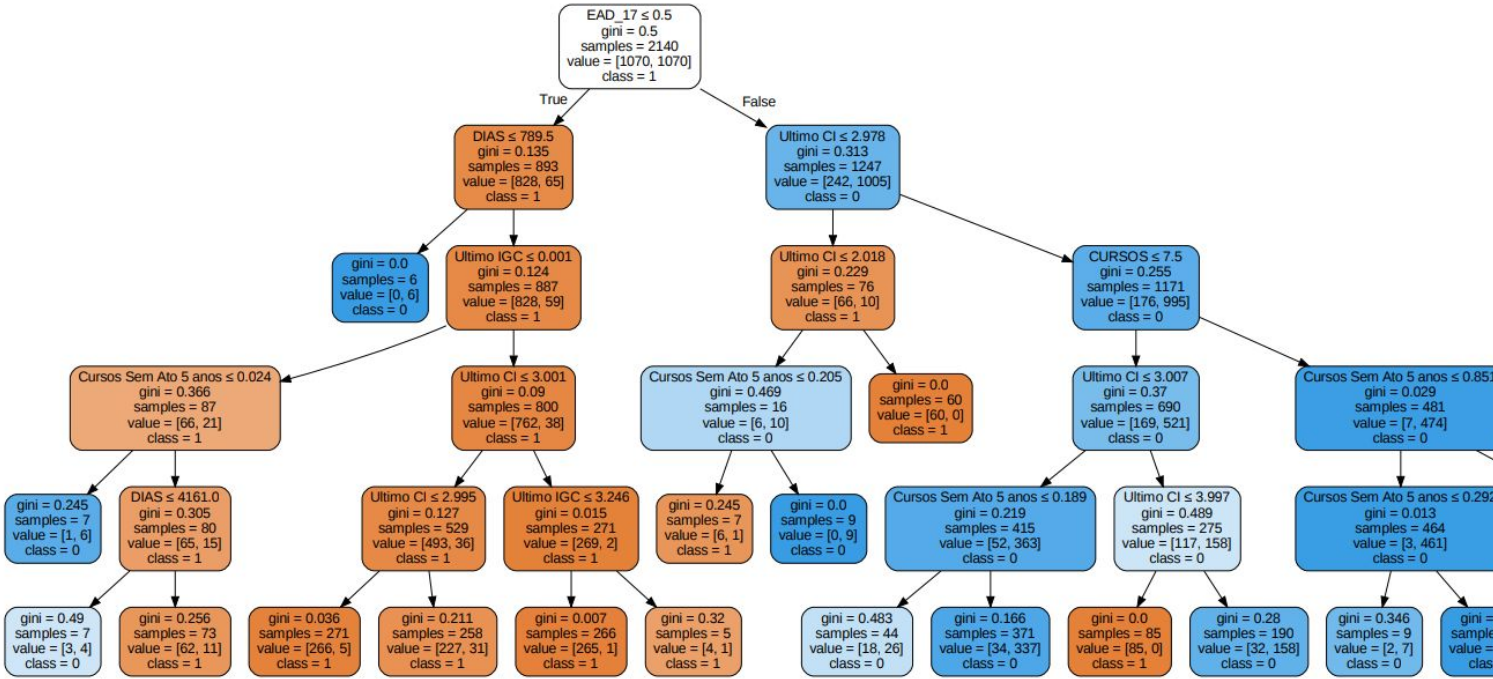
Métricas

	P	R	F1	S
Extinta	32%	58%	42%	19
Ativas	97%	91%	94%	265

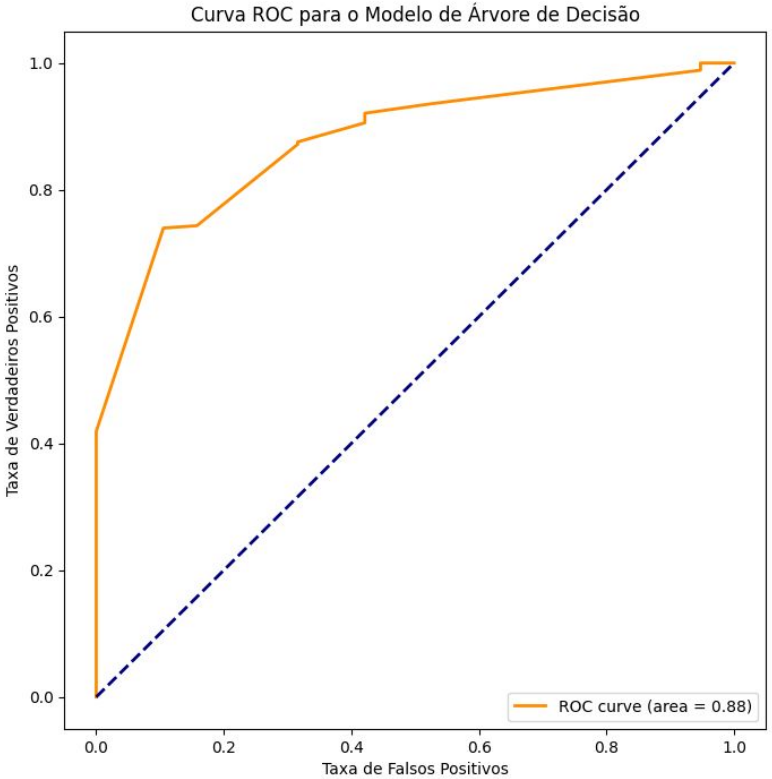
Acurácia: 89%



Visualização da árvore



Curva ROC


Curva ROC para o Modelo de Árvore de Decisão
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)

Modelo 3

RandomForest

Separeção dos Dados

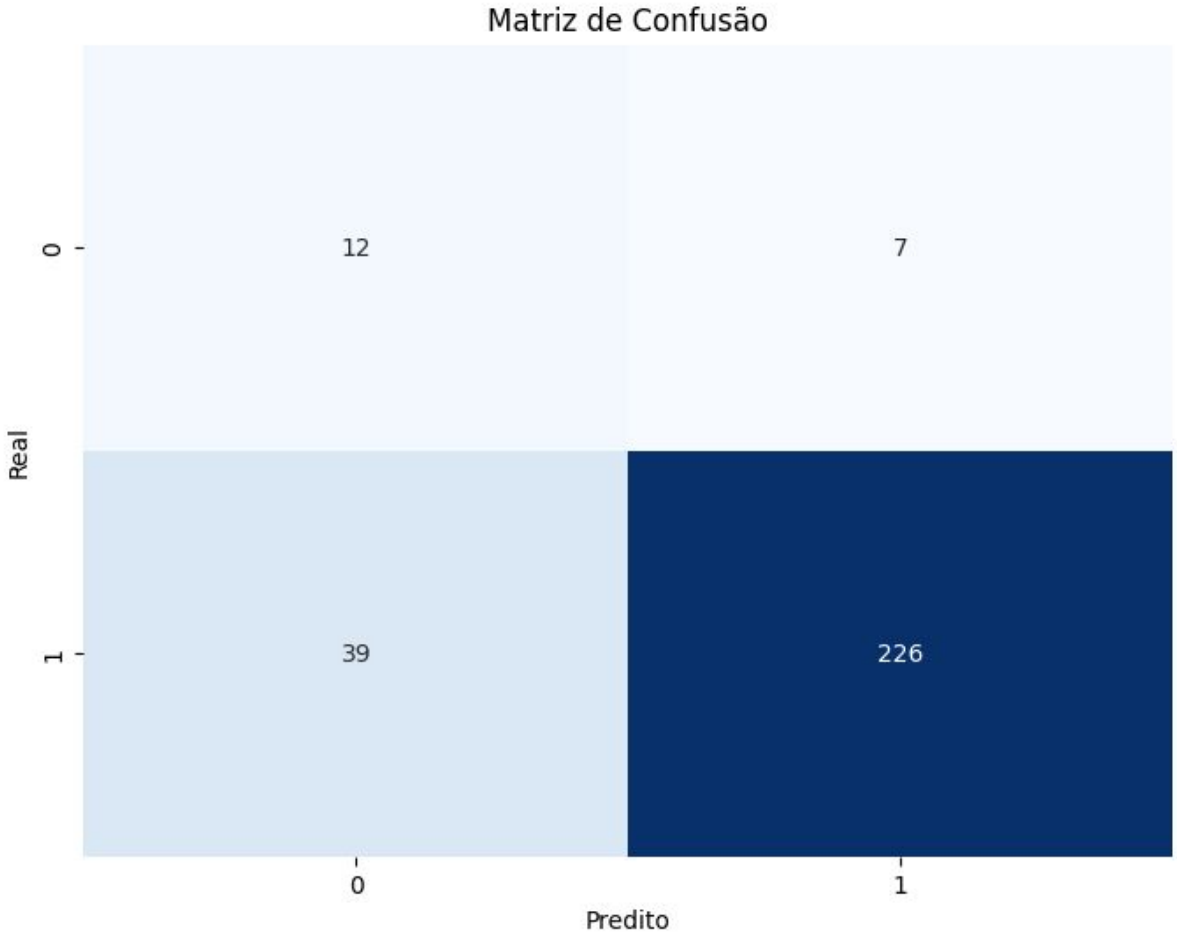
- 80% treino
- 20% teste

- n_neighbors = 3

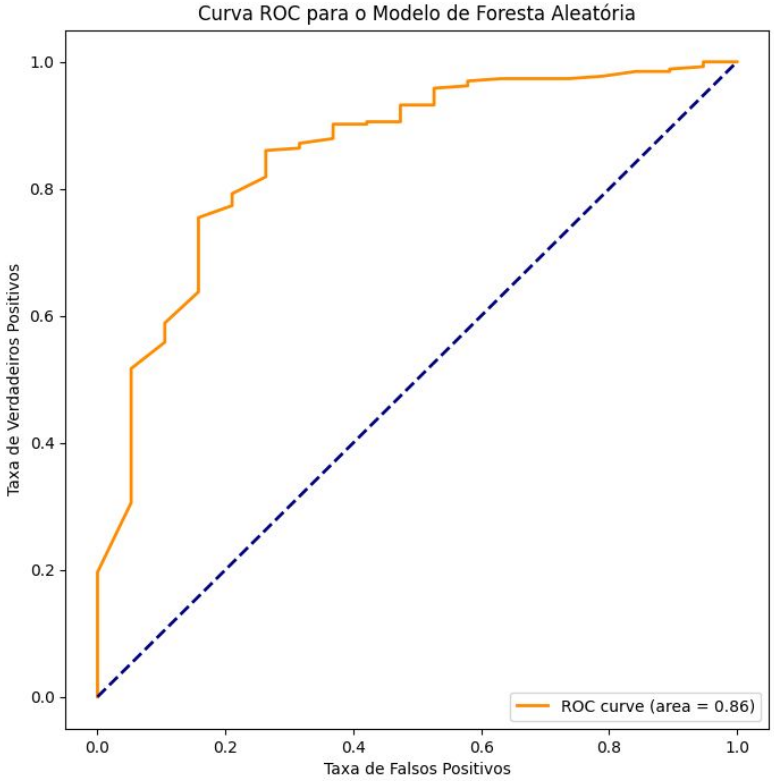
Métricas

	P	R	F1	S
Extinta	34%	53%	42%	19
Ativas	96%	93%	95%	265

Acurácia: 90%

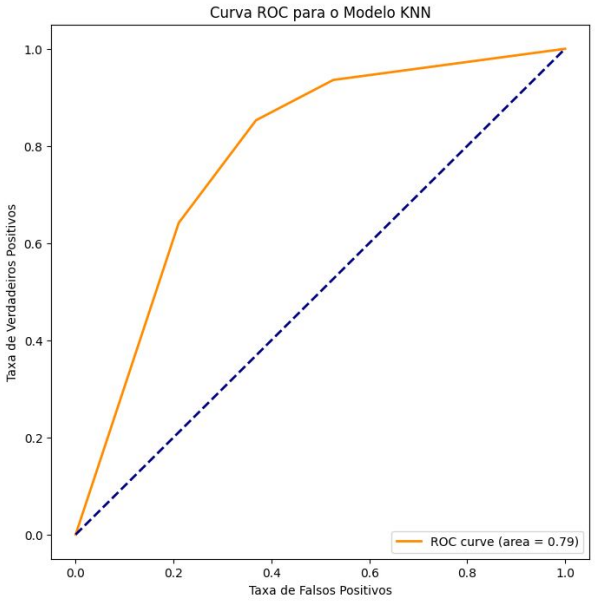


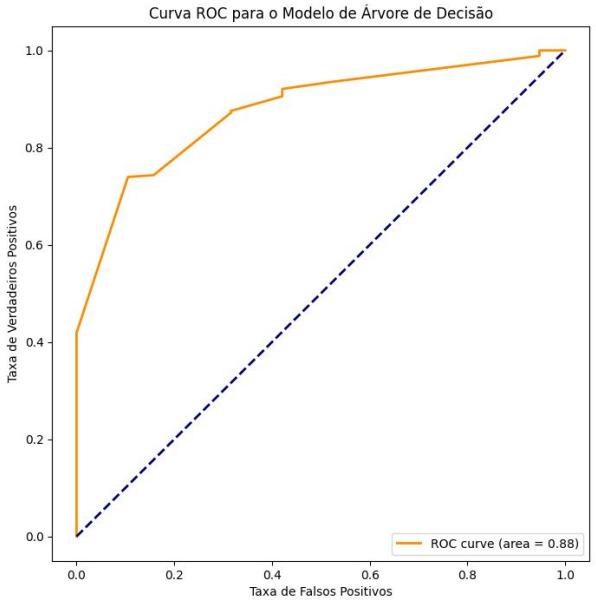
Curva ROC

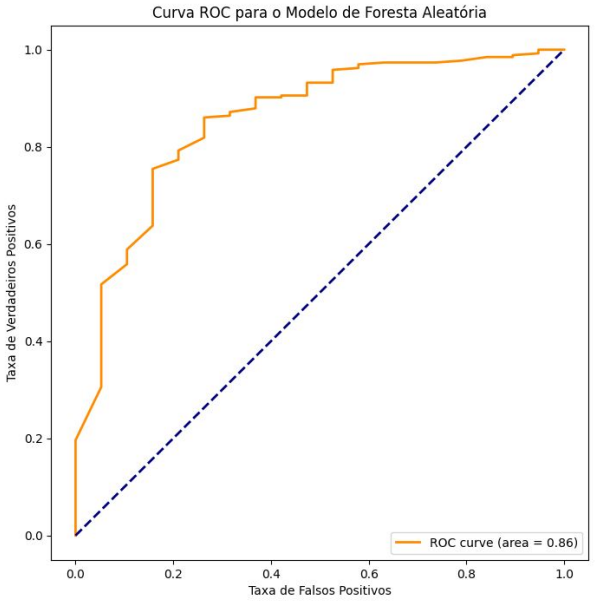

Curva ROC para o Modelo de Foresta Aleatória
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

Conclusões

Comparação da Curva







Comparação

	KNN	Árvore de Decisão	Floresta Aleatória
P(E)	24%	32%	34%
R(E)	63%	58%	53%
F1(E)	34%	42%	42%
P(A)	97%	97%	96%
R(A)	85%	91%	93%
F1(A)	91%	94%	95%
Acurácia	84%	89%	90%

Melhor Modelo

- Árvore de Decisão se mostrou mais interessante em relação a curva ROC

Principais Indicadores

- EAD_17
- Curso Sem Ato 5 anos
- DIAS
- Ultimo IGC
- Ultimo CI

Obrigado!